

Referee #1

Dear Referee #1, thank you so much for reviewing this paper and providing such a thorough and constructive review and comments. Here are our responses to your comments and suggestion point by points (in blue).

GENERAL COMMENTS

Air quality models are a vital tool for air quality research and management, but these models require the use of large input data sets and they generate even larger output data sets. Moreover, the size of these data sets is expected to grow with time, so the management and archival of such large data sets is an ongoing challenge for the air quality modeling community.

This paper describes a new and useful approach to this issue: an overall "lossy" compression algorithm that reduces the size of both input and output files while preserving their important features. This is done by combining a "lossy" precision-reduction conditioning of such data sets followed by lossless data compression. The paper describes this algorithm and then assesses the costs and benefits of this new data-set management approach in terms of disk space savings, model run times, and model accuracy. Both model-to-observations and model-to-model comparisons are considered for a 2016 annual simulation with the CMAQ air quality model along with a comparison of the efficacy of two well-known lossless data compression utilities.

I found this to be a well-structured and reasonably well-written paper that would be suitable for publication in GMD. I recommend its acceptance after a number of minor revisions, including some to improve its clarity. To this end I have made a number of specific comments and suggestions below that I believe will improve the final version and that I hope the authors will consider.

SPECIFIC COMMENTS

1. The paper considers both input emissions files and CMAQ output files but the title only references input files. Should the title be expanded slightly along the lines of "The Impact of Altering Emission and Output Data Precision on ..."?

Thank you so much for the suggestion. Although we did conduct an analysis on altered-precision CMAQ output, we consider the altered-precision emission data the 'driver' for this study. Due to its significance regarding CMAQ simulations (true emphasis of this study), we'll leave the title as is.

2. Terminology matters, and the manuscript is not always clear about exactly what is being done. The word "altered" is used a lot (32 times), but using a compound modifier like "altered-precision" or "reduced-precision" instead might be clearer. The terms "pre-processed" and "post-processed" are also used (lines 24, 25, 28, 105, 110, 117, 167) in what seems to me to be a confusing way, in that "pre-processing" is used in connection with emission files, which are input files upstream of the CMAQ model, whereas "post-processing" is used for CMAQ-generated output files. However, emission files are also output files that are generated by an emissions processing system like SMOKE, and for both types of

files the precision reduction is applied after the files have been generated. Again, referring to "reduced-precision emission files" and "reduced-precision CMAQ output files" might make it clearer that a transformation or conditioning step, namely precision reduction, has been applied to files after they have been generated.

We agree with the referee's suggestion. The manuscript has been modified accordingly. Since 'altered' was used numerous times throughout the manuscript, individual changes won't be listed here. Please examine the revised manuscript.

3. It would be helpful to the reader if a bit more detail were given in the Methodology section about (a) the characteristics of the emissions input files and CMAQ output files and (b) the 2016 base annual simulation ("orig"):

(a) The manuscript does not give any information about the temporal resolution of the emissions files or CMAQ output files -- do they contain hourly fields or more frequent or less frequent fields? How many different types of fields are there (e.g., different species, emissions vs. concentration vs. deposition fields)? What is the horizontal grid size? These details would help the reader to understand the size of the file sets.

The following sentences are added in the second paragraph of Section 2 Methodology, in the revised manuscript to address the referee's comments.

"For this study, CMAQ v5.3.1 (USEPA 2019, Appel et. al. 2021) was run with 459 columns, 299 rows, and 35 vertical layers with a horizontal grid-scale resolution of 12 km (Fig. 1.a). Emission input files consist of two area sources and nine point sources (hourly). The area source emission files contain 57 and 62 variables and the point source files contain anywhere from 54 to 58 variables (containing one vertical layer). Ten, CMAQ output files, nine of them are hourly, were generated in this study: Three output files were generated for simulation restart purposes (SOILOUT, CGRID which contains only one hour data, and MEDIA), two files contained average (APMDIAG and ACONC) and hourly (CONC) species concentrations, three files held wet deposition (WETDEP1; 140 variables), dry deposition (DRYDEP; 174 variables), and deposition velocity (DEPV; 104 variables) output, and lastly, the final file contained biogenic emission diagnostic output (B3GTS)."

(b) I know the details of the "orig" simulation are not directly relevant to the subject of this paper, but if the 'orig' simulation has already been documented in previous publications or reports, it would be helpful to have one sentence referring the reader to that additional information. For example, Table 6 states that the maximum daily PM2.5 bias was 512 ug/m³, a very large value, which left me wondering whether the 'orig' run included wildfire emissions (line 101 does mention "ptfire", but it is not clear whether this file was used in the 'orig' run).

The following reference has been added and cited in the second paragraph in section 2 Methodology.

Appel, K. W., Bash, J. O., Fahey, K. M., Foley, K. M., Gilliam, R. C., Hogrefe, C., Hutzell, W. T., Kang, D., Mathur, R., Murphy, B. N., Napelenok, S. L., Nolte, C. G., Pleim, J. E., Pouliot, G. A., Pye, H. O. T., Ran, L., Roselle, S. J., Sarwar, G., Schwede, D. B., Sidi, F. I., Spero, T. L., and Wong, D. C.: The Community

Multiscale Air Quality (CMAQ) model versions 5.3 and 5.3.1: system updates and evaluation, *Geosci. Model Dev.*, 14, 2867–2897, <https://doi.org/10.5194/gmd-14-2867-2021>, 2021.

This paper evaluated CMAQ version 5.3.1 in terms of model performance. We have adopted the same evaluation techniques from the referenced paper to determine the accuracy of our altered-precision simulations and cases. For the referenced paper (Appel et. Al, 2021), there were three kinds of fire emission input: ptfire (wildfire in US), ptagfire (agriculture related fire emission), and ptfire_othna (wildfire in neighboring countries). This is the cause for the relatively high, maximum daily PM2.5 bias (equal to 512 ug/m3).

4. References to daily measurements from the Ammonia Monitoring Network (AMoN) are incorrect. The sampling duration of the AMoN measurements is two weeks (e.g., <https://www3.epa.gov/castnet/docs/AMoNfactsheet.pdf>). This error should be corrected. It also raises concerns about whether the model predicted values were properly aligned with the observed values, that is, were they two-week averages. And since Tables 5 and 6 and Figure 4 present metrics for NH3 in units of $\mu\text{g}/\text{m}^3$, are the units of ppbv presented in Figures 5 and 8 for NH3 correct?

We thank the referee's thoroughness regarding the sampling duration of AMoN measurements. We used AMET to pair modeled output with respect to AMoN observations in time and space. This process was done automatically, by AMET, given a set of AMET arguments. We have spoken with an AMET author, Wyatt Appel, and got confirmation that AMET did pair modeled output to observations correctly according to AMoN measurement protocol.

For Table 5 and Table 6, they are model-to-observation comparisons, and the unit is $\mu\text{g}/\text{m}^3$. The model values were converted prior to pairing by AMET. For Figure 5 and Figure 8, they are model-to-model comparisons. The unit for CMAQ model output is ppmv (a simple conversion to ppbv). So we believe the units are correct.

The paper has been revised as follows:

Line 127 revised manuscript: "Likewise, daily averaged PM2.5 observations and two-week averaged NH3 observations were used to evaluate CMAQ."

All mentions of "daily NH3" have been replaced with "two-week averaged NH3."

5. Table 1 is used to illustrate the output of the precision-reduction algorithm. However, it does not fully describe how rounding is handled. For example, two lines could be added to show the A05, A04, and A03 values of 100150.0 and 100250.0.

Since 100150.0 and 100250.0 are similar in terms of rounding potential, we only used 100150.0 as an additional example in Table 1.

The following block of code is the gut of the reducing-precision algorithm.

```
write( fmt, '(a2, l2.2, a1, l2.2, a1)' '(e', ndigits+7, '.', ndigits, ')'  
  
read in input -> data
```

```
loop through c, r, and k
write(str,fmt) data(c, r, k)
read(str,*) datar(c,r,k)
```

write datar -> output

The fmt string determines the conversion of a real number into Fortran scientific notation with a specific number of digits after the decimal point. The converted number will be read in and then output to a netCDF file. Rounding follows regular standard (if the number you are rounding is followed by a number greater than 4, round the number up). The tool to reduce data precision is available at, DOI: 10.5281/zenodo.6620983, titled, "alter_data_src.tar.gz".

6. The definition of the statistical metrics on page 5, specifically on line 132, is not quite correct. The mean and standard deviations of the distribution are unknown, but they are estimated from the sample of model-measurement pairs, so the metrics are based on sample means and sample standard deviations.

Thank you for this correction. We have replaced "mean" and "standard deviation" with "sample mean of a distribution" and "sample standard deviation of a distribution", respectively.

Line 141-142 revised manuscript has been revised as "Where N is the total number of observed and predicted pairs, X is the observed value, Y is predicted value, σ is the sample standard deviation of a distribution, and the overbars in Eq. (2) refers to the sample mean of a distribution."

7. In the description of the Burrows-Wheeler algorithm, the clause "which chronologically reduces sequences of datasets by processing sequences through multiple layers of compression algorithms" (p. 5, l. 138) is not clear to me, especially the use of "chronologically". I know the Burrows-Wheeler transform is not easy to explain, but is some rewording or new wording possible?

We have revised the description of the Burrows-Wheeler algorithm in the manuscript as follows:

Line 147-148 revised manuscript: "In comparison, bzip2 uses the Burrows-Wheeler (Burrows and Wheeler, 1994) algorithm which sorts all possible rotations of an input lexically and forms an output by concatenating the last character from the sorted list."

8. There may be a discrepancy between the description of data set sizes in the Introduction (lines 57-58) and the discussion of data storage results in Section 3.1 (lines 159-165). If the CMAQ emissions files for one day are about 7 GB in size, how can the compression utilities reduce their size, again for a day, by 111 GB or 241 GB? Or if Section 3.1 is considering the emissions files for one entire year, then $7 \times 365 = 2,555$ GB and 21% and 48% of that number are 537 GB and 1,226 GB, respectively. Can some clarification be provided.

In lines 58 revised manuscript, it indicated a typical daily total file size of emission data, 7GB and it varies slightly each day. In lines 168-178 revised manuscript, it showed the daily relative reduction with respect to gzip and the total sum compressed file difference in the entire year of 2016. For example, the annual sum of the daily difference between compressed original emission data and compressed reduced-precision $n = 3$ emission data is about 111GB. The same explanation is true for bzip2 scenario. We have revised the manuscript to make it less confusing as follows:

“The gzip compression utility reduced the file sizes, on by an average by of 1 %, 5 %, and 21 %., This translates into about 5 GB, 26 GB, and 111 GB, actual difference between compressed orig case and relative reduction in file size of the compressed daily A05, A04, and A03 emissions datasets for the entire year of 2016, respectively, compared to the orig case. The reduction in file size (using gzip) was more substantial when applied to direct reduced-precision CMAQ output, with an average reduction in file size of 4 %, 19 %, and 647 %., This means or about 167 GB, 839 GB, and 2016 GB actual difference between orig case and for FXA05, FXA04, and FXA03, respectively for the entire year. With the bzip2 utility, the reduction in magnitude is much larger than with gzip, with an average reduction of file size equal to 6 %, 25 %, and 48 %, or (actual differences are about 27 GB, 126 GB, and 241 GB, respectively for A05, A04, and A03 emissions files and 19 %, 47 %, and 69 % (actual differences are about, or 856 GB, 2142 GB, and 3115 GB, respectively, for the compressed CMAQ output. Thus, bzip2 is found to be a more effective tool than gzip by roughly 5 %, 20 %, and 27 % for emission data and 15 %, 28 % and 23 % for CMAQ output, for A05, A04, and A03 reduced-precision by keeping 5, 4, and 3 significant digits (preprocessed reduced-precision emission and postprocessed reduced-precision output data), respectively.”

9. Section 3.2 would benefit from the addition of some discussion. Why should reducing the precision of the input emissions result in an overall reduction in run time but also cause increases in run times for some days? Why should there be a seasonal dependence in runtime differences, with larger differences for the first half of the annual simulation? Why are the run-time differences larger for the A05 and A03 simulations than for the A04 simulation? Are the decreases in overall run time expected and are they significant? Even if these questions cannot be answered with certainty, they should at least be raised and any possible explanations offered. The first paragraph of the Conclusion section (lines 263-264) also gives a questionable summary of Section 3.2, stating that "the A05, A04 and A03 simulations (were) consistently faster than the 'orig' simulation in a undedicated HPC system". This statement is contradicted by Figure 3, which shows run-time increases for at least some days of all three simulations run with precision-reduced emissions.

We do not have a satisfactory explanation of the relationship between reduced-precision data and run time shown in Fig 3. It was not one of the objectives of this study, however, this interesting behavior caught our eyes. We chose to report this in the paper as an observed outcome.

All the simulations were performed in an undedicated HPC system. During the execution of each case, the model competed for I/O resources with our simulation. Hence, I/O bottleneck could be to blame for the spike in relative run-time increase. Another potential explanation is that a change in emission input could alter the pathway in aerosol dynamics calculation or reduce the number of iterations in the chemistry solver. We have revised the manuscript slightly to reflect these two possible explanations.

The following addition was made at the end of Section 3.2 Runtime: “There are two possible explanations for such behavior: First, during the execution of each case, CMAQ competed for I/O resources with other tasks on the system. As a result, an I/O bottleneck could explain spikes in relative run-time on certain simulation days (Fig. 3). Second, a change in emission input (due to the reduced-precision emission data) could alter the pathway for the aerosol dynamics calculation. This change in emission input can also reduce the number of iterations in the chemistry solver.”

We used the word “consistently” to describe the overall, or average, behavior of runtime. We will replace “consistently” with “for most days” in the manuscript to avoid confusion.

10. In Section 3.3 a comparison is made between the ranges of maximum absolute bias, that is, an analysis of model errors, and the concentration values of the air quality standards (lines 202-203) but the reason for this comparison was unclear to me? Second, in the methodology description in line 206 was RMSE really first calculated for all hourly grid-grid pairs in 2016, or do you mean that grid-grid pair differences were first calculated? Otherwise, it is not clear how average hourly RMSE could then be calculated for each season and region. And when you state "all grid-grid pairs", do these include grid cells located outside the contiguous U.S.? And in lines 209-211 where you state that "total accumulative RMSE for PM2.5, O3, and NH3 (sum of all region's RMSE) did not exceed 0.1 $\mu\text{g}/\text{m}^3$, 0.4 ppbV, and 0.1 ppbV, respectively for all cases and for all seasons", based on Figure 5 could these values not be smaller (perhaps 0.04 $\mu\text{g}/\text{m}^3$, 0.3 ppbV, and 0.05 ppbV)?

We will address this in three parts.

Part A: It is intended to show how insignificant the differences are with respect to current air quality thresholds. Since the values are noticeably small, we will remove this line to reduce the inclusion of unnecessary information.

The following sentences were removed, “In contrast, the 24-hour fine particle limit for PM2.5 exposure is 35 $\mu\text{g m}^{-3}$ and the 8-hour exposure limit for O3 is 70 ppb (U.S. EPA Criteria Pollutants, <https://www.epa.gov/criteria-air-pollutants/naaqs-table>). Thus, the ranges of maximum absolute bias for 24-h PM2.5 and MDA8 O3 do not exceed air quality criteria ($0.16 \mu\text{g m}^{-3} < 35 \mu\text{g m}^{-3}$) and $0.13 \text{ ppb} < 70 \text{ ppb}$) which suggests the differences in simulations and cases are negligible.”

Part B: RMSE was calculated for each hour using all model-model pairs (at one hour) within a mask region which are illustrated in Fig. 1.a. Pairing was only done for cells within shaded regions over the contiguous US.

The following alteration was made to lines 218-219 of the revised manuscript: “Only cells that fell within each region (Fig. 1.a), within the contiguous US, were used to calculate hourly RMSE for all available regional pairs.”

Part C: The values 0.1 $\mu\text{g}/\text{m}^3$, 0.4 ppbV, and 0.1 ppbV were chosen due to the plotted maximum (Figure 9.) y-tick values.

The following alteration was done to lines 226-227 of the manuscript: “For example, the total accumulative RMSE for PM2.5, O3, and NH3 (sum of all region's RMSE) did not exceed 0.04 $\mu\text{g m}^{-3}$, 0.3 ppbV, and 0.05 ppbV, respectively for all cases and for all seasons.”

The following alteration was made to lines 238-240 of the manuscript: "(Fig. 5. Total accumulated values did not exceed 0.04 $\mu\text{g m}^{-3}$), 0.3 ppbV, and 0.05 ppbV for PM_{2.5}, O₃, and NH₃ respectively)"

11. The analysis of accuracy for hourly deposition rates at the end of Section 3.3 is helpful, but it does not directly address the main impacts of deposition, which are cumulative. Could something be said about the accuracy of seasonal or annual deposition values for the simulations and cases?

Unfortunately, we only evaluated hourly deposition rates for this study (annual). The last paragraph in Section 3.3 has been revised as:

"The final aspect of this evaluation explores aspects of important species using boxplots (minimum, 25th percentile, median, 75th percentile, and maximum) of hourly, spatially averaged r , BIAS, NMB, and RMSE for all grid-to-grid pairs (across CMAQ's domain) for deposition rates with respect to the orig case. Boxplots were created for the wet deposition rates of sodium (Na), ammonium (NH₄), chlorine (Cl), nitrate (NO₃), sulfate (SO₄), and the dry deposition rate of O₃ for all altered-precision simulations and cases. For all deposition rates, the altered-precision ndigit 3 cases (FX0503, FX0403, and FX0303) performed worst, relatively speaking, with respect to the original simulation despite negligible RMSE error differences. The altered-precision ndigit 4 cases (FX0504, FX0404, and FX0304) performed nearly identically to the ndigit 5 cases for all deposition rates excluding the wet-deposition rate of sodium and the dry deposition rate of ozone (a relatively small increase in error). The altered-precision 5 cases (FX0505, FX0405, and FX0305) and the altered simulations (A05, A04, and A03) performed nearly identically to the original simulation for all deposition rates. Interestingly, amongst the altered-precision simulations (A05, A04, and A03), A05 performed worst (very negligible error) with respect to the original simulation for all deposition rates, excluding the wet deposition rate of chlorine, when considering NMB, MB, and r . Overall, all species, simulations, and cases performed similarly with respect to the orig case, and hence, amongst each other (Fig. 9)."

TECHNICAL CORRECTIONS/SUGGESTIONS

p. 1, l. 13: Perhaps "... archive input and output data sets" or "... archive input and output files"

Done (chose the latter one).

p. 1, l. 16: "desired post-processing of the output (e.g., for evaluations or graphics)"?

Done.

p. 1, l. 21, 23: Change "losslessness compression" to "lossless compression"

Done.

p. 1, l. 22: Don't you mean "before" rather than "after"?

It should be "after". The new approach consists of two parts: 1. Reduce data precision in a file, 2. Apply lossless compression to the reduced-precision dataset. This study reports by reducing the data precision

enhances the data compression rate. In other words, reducing data precision in a file does not reduce storage but after a lossless compression utility is applied.

p. 1, l. 27: Would this be more accurate: "To enhance the analysis of disk space efficiency, the output from the altered emissions CMAQ simulations ..."?

Done (we have also adapted the suggestion from specific comment #2).

p. 1, l. 30: Would this be more clear: "Thus, in total, 13 gridded output products (four simulations and nine altered output cases) were ..."? (cf. line 108)

Done with the adaption from specific comment #2.

p. 3, l. 70: Perhaps "... by manipulating the mantissa of individual floating-point numbers".

Done.

p. 3, l. 90: "This study proceeds as follows:"

Done.

p. 4, l. 105: "Emission input and CMAQ output data were then compressed" -- this addition would emphasize that the lossy algorithm discussed in this manuscript consists of two steps.

Agreed, but changed to: "Emission input and CMAQ output data were compressed separately ..."

p. 4, l. 110: Perhaps "... (see Table 2 for a full list of simulations and cases)"

Done.

p. 5, l. 142: Perhaps "Examples of precision-reducing transformation of floating-point numbers from their ..."

Done.

p. 7, l. 157: Perhaps "... throughout the entirety of the 2016 simulation"

Done.

p. 7, l. 161: Perhaps "... when applied to reduced-precision CMAQ output files"

Done.

p. 7, l. 162: Should this be "FX05, FX04, and FX03"?

Done. We have revised the last sentence in section 3.1 Data Storage for clarity as "... output, for A05, A04, and A03 ..." to "... output, for reduced-precision by keeping 5, 4, and 3 significant digits ..."

p. 9, l. 183: Perhaps "... all available model-measurement pairs throughout 2016".

Done.

p. 10, l. 20: Perhaps "... maximum absolute bias for 24-h PM2.5 and MDA8 O3 do not ...".

We believe l. 20 refers to l. 203. Done.

p. 19, l. 269: Perhaps change "mimic" to "are similar to those for".

Done.

p. 19, l. 281, 285: Rather than "performance" and "quality", do you really mean "accuracy"?

Those two words have been replaced with "accuracy".

Fig. 4 and 5 captions: Perhaps expand "(x)" and "(y)" to "(x-axis)" and "(y-axis)".

Done.

Table 7 caption: Perhaps "... with respect to the 'orig' simulation across all grid cells".

Done.